La aceleración de la navegación web en los institutos

Diego Martín Arce[1] - Juan Sanguino González[2]

[1]Director del IES Calamonte - diegomartina@edu.juntaextremadura.net

[1]Profesor de Geografía e Hª. IES Calamonte - juan.sanguino@edu.juntaextremadura.net

INTRODUCCIÓN

La configuración de los institutos extremeños, con la utilización masiva de ordenadores en el proceso de enseñanza aprendizaje, plantea numerosos retos. Uno de los problemas sobre los que existe consenso es la velocidad de los accesos a internet, en muchos casos, dependiendo de la hora, más lenta que una conexión telefónica normal. Aunque el ancho de banda puede parecer suficiente (2Mb) la conexión de doscientos-trescientos equipos ralentiza extraordinariamente la navegación. Naturalmente esto hace poco útil plantearse el acceso a webs externos en la enseñanza cotidiana. Este problema puede solventarse en gran medida si utilizamos una de las herramientas menos conocidas y que se incluye en la distribución de gnuLinEx: wget. Podemos utilizar wget para descargar webs enteras, o partes de ellas, y luego utilizarla con nuestros alumnos, prescindiendo de los problemas de velocidad de la red, pues al descargarlas podemos navegar por esas páginas a la velocidad de nuestra red LAN: 100Mb por segundo. Desde luego la consulta a estas webs debe estar prevista, pues las páginas necesitan ser descargadas con anterioridad. Esta herramienta no sirve cuando de lo que se trate sea de buscar información. Aunque también sirve para descargas mediante ftp este tema no será abordado en este artículo por razones de extensión.

1. INSTALACIÓN Y CARACTERÍSTICAS

WGET es una completa utilidad gnu/LinEx que trabaja en modo texto y que sirve para bajar ficheros usando los protocolos HTTP y FTP.

(1) Hay que señalar que no es necesario instalar esta aplicación, porque ya está instalada en la distribución de gnuLinEx.

(2) En segundo lugar, es una herramienta utilizable, desde un terminal. No se utiliza ninguna interfaz gráfica.

(3) La sintaxis es sencilla y, como casi todo en gnuLinEx, muy configurable.

(4) Por último, tiene numerosas opciones, muchas de las cuales no vamos a utilizar nunca o casi nunca.

Dado el carácter práctico de este artículo, es recomendable, leerlo usando un ordenador conectado a Internet para poder seguir los ejemplos.

2. MANEJO DE LAS OPCIONES

La utilización de wget puede desglosarse en tres partes: comando, opciones y URL. Aunque puede usarse sin ninguna opción.

2.1. Opciones básicas

En este apartado vamos a analizar la utilidad de tres opciones: -r -k y -l

Veamos un primer ejemplo:

```
$wget http://www.rte-extremadura.org
```

Si ahora abrimos un navegador y le indicamos en la barra de navegación la ruta de el archivo index.html podemos ver que tenemos esa página en nuestro disco duro y que se ha descargado a una gran velocidad.

A continuación vamos a ver hacer lo mismo pero con la opción -r:

```
$wget -r http://www.rte-extremadura.org
```

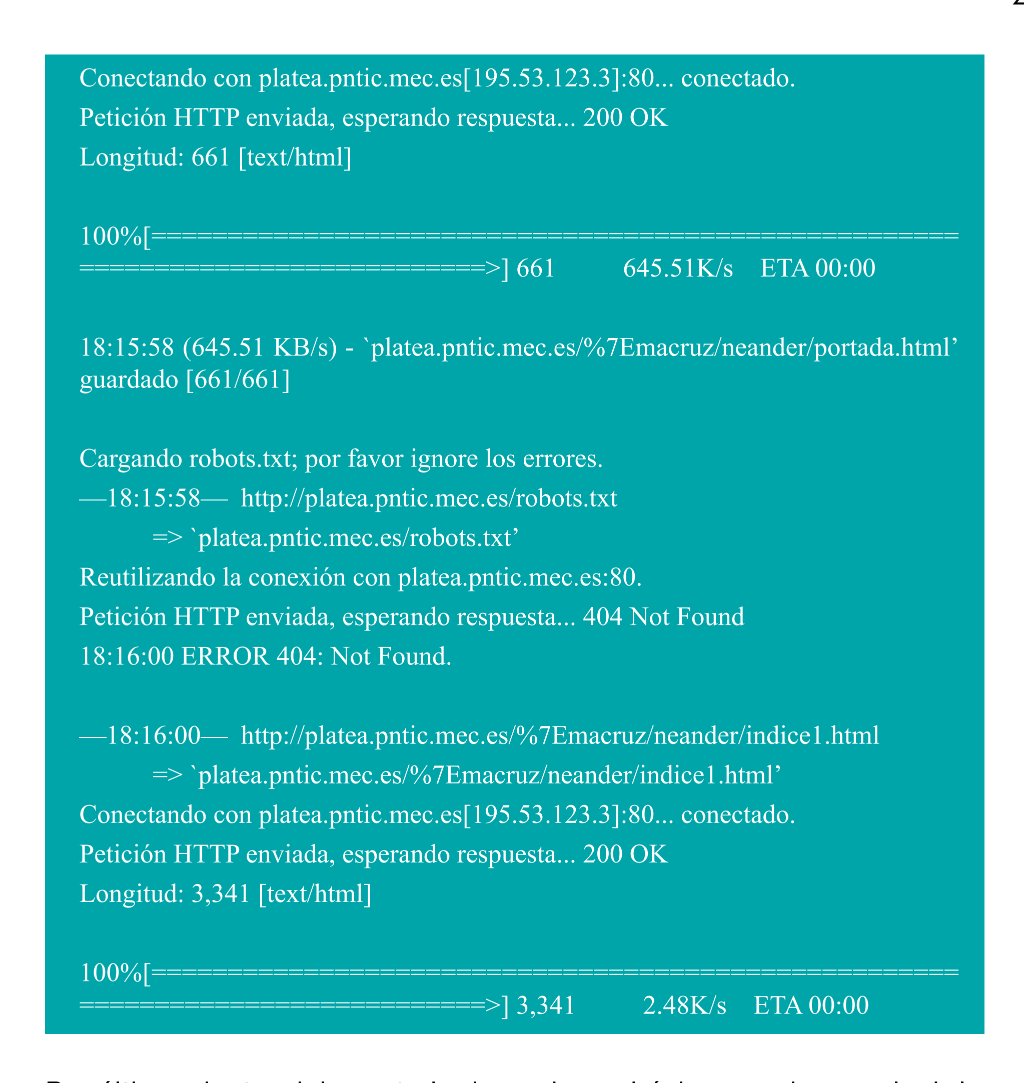
Ahora observamos que no se detiene en una sola página sino que descarga muchas. Si nos interesa podemos dejar que descargue enteramente la web www.extremadurasi.org, si no podemos cortar la descarga pulsando simultáneamente Ctrl+z.

Si ahora abrimos el navegador vemos que wget nos ha creado una carpeta denomina con el nombre de la URL en nuestro caso www.rte-extremadura.org y abrimos el archivo index.html vemos que tenemos todas las páginas pero que los enlaces apuntan fuera de nuestro web. Para lograrlo tenemos que usar la opción -k que sirve para transformar los enlaces en locales.

```
$wget -r -k http://www.rte-extremadura.org
```

Obtenemos en el terminal una serie de mensajes parecidos a éstos

```
sergio@sergio:~$ wget -r -k -l3
http://platea.pntic.mec.es/~macruz/neander/portada.html &
—18:15:57—  http://platea.pntic.mec.es/%7Emacruz/neander/portada.html
           => `platea.pntic.mec.es/%7Emacruz/neander/portada.html'
[1] 2141
sergio@sergio:~$ Resolviendo platea.pntic.mec.es... hecho.
```



```
Conectando con platea.pntic.mec.es[195.53.123.3]:80... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 661 [text/html]

100%[==========================================================
============================>] 661          645.51K/s    ETA 00:00

18:15:58 (645.51 KB/s) - `platea.pntic.mec.es/%7Emacruz/neander/portada.html'
guardado [661/661]

Cargando robots.txt; por favor ignore los errores.
—18:15:58—  http://platea.pntic.mec.es/robots.txt
           => `platea.pntic.mec.es/robots.txt'
Reutilizando la conexión con platea.pntic.mec.es:80.
Petición HTTP enviada, esperando respuesta... 404 Not Found
18:16:00 ERROR 404: Not Found.

—18:16:00—  http://platea.pntic.mec.es/%7Emacruz/neander/indice1.html
           => `platea.pntic.mec.es/%7Emacruz/neander/indice1.html'
Conectando con platea.pntic.mec.es[195.53.123.3]:80... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 3,341 [text/html]

100%[==========================================================
============================>] 3,341          2.48K/s    ETA 00:00
```

Por último, dentro del apartado de opciones básicas, podemos elegir la "profundidad" de descarga que queremos con la opción -l seguida de un número (si no indicamos nada, wget toma la opción por defecto 5 niveles de profundidad).

Borremos antes de continuar el directorio donde se encuentra nuestra descarga anterior para que no nos lleve a confusión. Y a continuación procedemos como sigue:

```
$wget -r -k -l2 http://www.rte-extremadura.org
```

Hasta ahora hemos visto cómo utiliza wget para descargar páginas individuales, cómo se utiliza para descargar páginas recursivamente y cómo seleccionamos la profundidad de la descarga.

Estas tres opciones podemos llamarlas básicas y conviene que practiquemos algo más con ellas por nuestra cuenta. A continuación vamos a estudiar opciones muy útiles, pero no imprescindibles.

2.2. Opciones avanzadas.

En esta sección vamos a descubrir algunas opciones que nos van a servir para perfilar la descarga de datos de las direcciones que queramos descargar.

2.2.1. -p (page-requisites)

La estructura de las páginas HTML consiste en una serie de ficheros aislados que son llamados desde otros. En el caso del fichero index.html que tenemos alojado en nuestro localhost vemos que aparece una imagen con un ñu y un pingüino. Esa imagen es un fichero distinto que es llamado desde el fichero index.html. Al igual que con este ejemplo ocurre si la página incluye música, vídeo, un CSS, etc... Para garantizarnos la descarga de la página con todos los elementos necesarios para su contemplación tenemos que recurrir a la opción -p.

2.2.2. -c (continue)

Con esta opción se reanuda la descarga de un fichero parcialmente descargado, por ejemplo por un corte de electricidad o porque hemos suspendido la descarga voluntariamente por cualquier circunstancia. Si una descarga ha sido truncada, podemos iniciar otra y la retomará desde el punto en el que la dejó.

2.2.3. -i (spider)

Como su nombre indica actúa como un spider. No baja los ficheros sino que chequea que están allí. Es útil para la gestión de marcadores de página (bookmarks).

2.2.4. -E (extension)

Añade la extensión .html al fichero descargado y convertido en local. Sirve para hacer un espejo de un sitio remoto que usa páginas .asp y que se quieren hacer visibles. Otro uso es cuando se quiere descargar la salida de los CGI. Actuando de este modo cada cierto tiempo se vuelve a ese fichero. Para prevenir esta descarga inútil se debe usar -k y -K para que la versión original del fichero sea salvada como X.orig.

2.2.5. -http-user=USUARIO y -http-password=CONTRASEÑA

Estas opciones son necesarias para descargarnos páginas en las que nos van a pedir un usuario y una contraseña. Hay que destacar que si ejecutamos esta opción estamos dejando expuesto nuestro login y contraseña para acceder a esa web.

```
$wget -r -k -http-user=escandinaviA —http-password=frio  http://loquesea.com
```

2.2.5. —random-wait

Algunas web hacen un análisis de registros (*logs*) para buscar similaridades estadísticas significativas en el tiempo de petición de las páginas. Con esta opción se enmascara la presencia de wget para los análisis de estos análisis.

2.2.6. -A (accept-list)

Esta opción permite elegir listas de archivos que contengan en su nombre una cadena de caracteres. Por ejemplo, la opción -A .gif descargaría sólo los ficheros de extensión gif. Pueden seleccionarse varias extensiones separadas por comas: -A gif, jpg. En cambio, si ponemos -A fich* nos descargará todos los ficheros que comiencen por la cadena “fich”.

2.2.7. -R (reject-list)

Hace justo lo contrario que la opción -A. No serán descargados los ficheros que contengan las cadenas a las que se haga referencia.

2.2.8. -K (backup)

Cuando convierte un archivo hace una copia de respaldo de la versión original con la extensión .orig

2.2.9. -P (Prefix)

Con esta opción le indicamos dónde queremos que guarde los ficheros que baje (directorio prefijado). Si no se indica nada los guarda en el directorio actual. Hay que tener cuidado para no confundirla con la opción -p

Ejemplo:

```
$wget -p -P/home/linex/datos http://www.rte-extremadura.org
```

De este modo se guarda la primera página de la web
http.//www.extremadurasi.org en /home/linex/datos

2.2.10. -m (mirror)

Cuando se selecciona activa las opciones de recursivo, marcador de tiempo, profundidad y recursividad infinita y permanecen los directorios de listados de FTP. Es el equivalente a marcar a la vez -r -l -K inf -rn.

2.2.11. -np (no parent)

Con esta instrucción indicamos a wget que no ascienda en el árbol de jerarquía de la web que estemos descargando. Es muy útil si lo que queremos es descargar varias páginas, pero no nos interesan las anteriores. Como no sabemos cómo han construido la web, conviene que usemos junto a esta opción -p.

Con estas opciones tenemos más que suficiente para descargar webs que nos sirvan para acelerar la navegación web en las aulas.

2.1.12. -nd (no-directory)

Esta opción se usa cuando no queremos que recree la jerarquía de directorios porque no nos va a ser útil (p.e.: queremos guardar todos los ficheros de un tipo en un solo directorio)

Para descubrir más opciones y su significado puede acudirse al manual de wget (man wget), desde un terminal.

Una pequeña ayuda y su salida la obtenemos si tecleamos

```
sergio@sergio:~$ wget -help
GNU Wget 1.8.1, un recuperador por red no interactivo.
Modo de empleo: wget [OPCIÓN]... [URL]...
Los argumentos obligatorios para las opciones largas son también obligatorios
para las opciones cortas.

Inicio:
  -V, —version              muestra la versión de wget y termina.
  -h, —help                 muestra esta ayuda.
  -b, —background           pasa a segundo plano al iniciar.
  -e, —execute=ORDEN        ejecuta una orden como las de `.wgetrc'.

Fichero de entrada y registro:
  -o, —output-file=FICHERO       registra los mensajes en FICHERO.
  -a, —append-output=FICHERO     añade los mensajes a FICHERO.
  -d, —debug                     imprime la salida de depurado.
  -q, —quiet                     modo silencioso (no muestra ninguna salida).
  -v, —verbose                   modo informativo (predeterminado).
  -nv, —non-verbose              muestra el mínimo necesario de información.
  -i, —input-file=FICHERO        descarga las URLs que haya en FICHERO.
  -F, —force-html                trata el fichero de entrada como HTML.
  -B, —base=URL                  añade URL delante de los enlaces relativos
                                 en el fichero -F -i.
      —sslcertfile=FICHERO       certificado opcional del cliente.
      —sslcertkey=FICHERO        llave opcional para este certificado.
      —egd-file=FICHERO          fichero del socket EGD.

Descarga:
      —bind-address=DIRECCIÓN    realiza un bind a la DIRECCIÓN (máquina o IP)
                                 en la máquina local.
```

-t, —tries=NÚMERO	establece en NÚMERO el número de reintentos (0 no pone límite).
-O, —output-document=FICHERO	escribe los documentos en FICHERO.
-nc, —no-clobber	no sobrescribir ficheros existentes. o utilizar sufijos .#
-c, —continue	continuar recuperando un fichero existente.
—dot-style=ESTILO	establece el estilo de la pantalla de recuperación
-N, —timestamping	no recupera ficheros más viejos que los locales.
-S, —server-response	imprime la respuesta del servidor.
—spider	no recupera nada.
-T, —timeout=SEGUNDOS	establece el tiempo de espera de lectura en SEGUNDOS.
-w, —wait=SEGUNDOS	espera SEGUNDOS entre recuperaciones.
—waitretry=SEGUNDOS	espera 1...SEGUNDOS entre reintentos.
—random-wait	espera de 0 a 2*WAIT segundos entre reintentos.
-Y, —proxy=on/off	habilita/deshabilita el uso de proxies.
-Q, —quota=NÚMERO	establece la cuota de recuperación en NÚMERO.
—limit-rate=TASA	limita la tasa de descarga a TASA.

Directorios:

-nd —no-directories	no crea directorios.
-x —force-directories	fuerza la creación de directorios.
-nH, —no-host-directories	no crea directorios en el anfitrión
-P, —directory-prefix=PREFIJO	guarda ficheros en PREFIJO/...
—cut-dirs=NÚMERO	descarta NÚMERO componentes del directorio remoto.

Opciones de HTTP:

—http-user=USUARIO	establece que el usuario de http es USUARIO.
—http-passwd=CLAVE	utiliza CLAVE como contraseña de http.
-C, —cache=on/off	(des)habilita la caché del servidor de datos. (normalmente habilitada).
-E, —html-extension	guarda todos los ficheros de texto/html con la extensión .html.
—ignore-length	ignora el campo ‘Content-Length’ de la cabecera.

—header=TEXTO	inserta el TEXTO entre las cabeceras.
—proxy-user=USUARIO	establece que el usuario del proxy es USUARIO.
—proxy-passwd=CLAVE	utiliza CLAVE como contraseña del proxy.
—referer=URL	incluir cabecera ‘Referer: URL’ en petición HTTP.
-s, —save-headers	guarda las cabeceras de HTTP en un fichero.
-U, —user-agent=AGENTE	identificarse como AGENTE en vez de Wget/VERSIÓN.
—no-http-keep-alive	deshabilita las conexiones persistentes de HTTP.
—cookies=off	no utiliza cookies.
—load-cookies=FICH.	carga las cookies desde FICH. antes de la sesión.
—save-cookies=FICH.	guarda las cookies en FICH. tras la sesión.

Las opciones por defecto que generan estos parámetros pueden modificarse copiando el fichero /etc/wgetrc como /home/usuario/.wgetrc.

3. CONSTRUCCIÓN DE UN PROYECTO.

Una vez analizadas las principales opciones vamos a desarrollar un proyecto, que vamos a dividir en dos subproyectos. En el primer subproyecto vamos a descargar páginas de dos webs y a hacerlas navegables dentro de la red local. En el segundo haremos lo mismo pero automatizando la descarga de modo que nos encontremos las páginas descargadas cuando lleguemos al trabajo. Como ejemplo para este segundo caso crearemos un quiosco electrónico, para lo cual necesitaremos hacer uso de la tabla del cron (crontab). Una vez que tengamos la suficiente soltura no consumiremos más de una hora en realizarlo todo, teniendo en cuenta que este periodo puede superarse dependiendo de la velocidad de conexión.

(1) Instalación de un servidor web

Para instalar un servidor web en un equipo sólo hay que descargar de los repositorios Debian o LinEx dos paquetes: apache y apache-common. Puede utilizarse synaptic (administrador de paquetes) o escribir en una terminal como root:

```
#apt-get install apache apache-common
```

Una vez descargados ese equipo contará con un servidor web. Para comprobarlo abrimos un navegador (Mozilla, Galeón o cualquier otro) y escribimos en la barra de navegación http://localhost/ con lo que nos aparecerá una página como esta:



Fig. 1. *La página de bienvenida de nuestro servidor web-apache.*

(2) Elegir los recursos y usar wget.

Hemos elegido dos webs para descargar sus archivos. La primera es una página chilena de Geografía, mientras que la segunda es una página de Prehistoria.

En el primer caso vamos a descargar la página de manera recursiva (-r) transformando en local todos sus enlaces (-k) y con tres niveles de profundidad (-l3). Le decimos, además que descargue todo lo necesario para que pueda verse la página (-p) y el lugar donde queremos que guarde la información en /var/www (-P/var/www).

```
$wget -r -k  -p -P/var/www http://icarito.tercera.cl/icarito/2001/831/
```

En el segundo caso vamos a descargar una web con contenidos sobre evolución humana manera recursiva (-r), local (-k) y guardando la información que genere en el /var/www (-P/var/www).

```
$wget -r -k -p -P/var/www http://www.ucm.es/info/paleo/ata/port-nt.htm
```

En tercer lugar vamos a descargar imágenes de tipo jpg (-A .jpg) de modo recursivo (-r) de una web que podremos usar en nuestras clases para hacer presentaciones con Impress. Queremos además que nos las guarde en una carpeta en /home/linex/Documentos/imágenes

(-P/home/linex/Documentos/imágenes) sin recrear los directorios originales

(-nd). Vamos a tomar una web con imágenes de escultura italiana desde la Edad Media hasta la Edad Moderna.

```
$wget -r -nd -A .jpg -P/home/linex/Documentos/imágenes
http://www.thais.it/scultura/default.htm
```

(3) Construcción de un quiosco electrónico.

Ahora construimos un fichero índice que nos va a llevar a las distintas secciones del quiosco y lo guardamos con el nombre index.html en el directorio /var/www, sustituyendo al fichero anterior. En este fichero crearemos los enlaces a los ficheros índice de las publicaciones que hayamos seleccionado. Para ello podemos utilizar el programa de construcción de páginas web de Mozilla (Composer/Medellín).

Las direcciones URL que vamos a necesitar son:

Para el diario Hoy.
http://www.hoy.es/
Para la revista National Geographic
http://www.esmas.com/nationalgeographic/
Para la revista Muy Interesante
http://www.muyinteresante.es/

Ahora editamos el fichero /etc/crontab y añadimos las siguientes líneas

```
#Bloque del quiosco
#diario hoy
35 16 * * *     root    wget -r -k -p -l4 -np -P/home/sergio/quiosco
http://www.hoy.es
#National Geographic en español
27 6 8 * *     root    wget -r -k -p -l3 -np -P/home/sergio/quiosco
http://www.esmas.com/nationalgeographic/
#Muy interesante
30 6 10 * *     root    wget -r -k -p -l4 -np -P/home/sergio/quiosco
http://www.muyinteresante.es/
#Fin del bloque de quiosco
```

Nos llamará la atención el principio de cada línea en las que hay una serie de números y después asteriscos. Esto hace referencia a la hora y los días. Veamos el caso del National Geographic

27 6 8 * *

Quiere decir que a las 6 horas 27 minutos del día 8 empezará la descarga. Los asteriscos indican que serán todos los meses y cualquier día de la semana.



Fig. 2. *La página inicial concluida y con los enlaces a las descargas.*

CONCLUSIONES

wget es una potente herramienta para resolver la lentitud de la navegación web en horas de clase. Esta herramienta permite el acceso y descarga de las páginas de Internet acelerando extraordinariamente su navegación. Por otro lado, al consumir sólo recursos internos deja más banda para aquellos usuarios que necesiten salir de la red local.

REFERENCIAS

Para conocer más opciones de wget conviene leer la página del manual correspondiente mediante la orden

```
$wget wget
```